Automatic authorship attribution based on character n-grams in Swiss German

Oppliger, Rahel

Abstract: Automatic authorship attribution aims to train computers to identify the author of a disputed text based on idiolectal language features. When confronted with nonstandard data – in the present study Swiss German instant messages – languagespecific NLP toolkits are often unavailable, limiting the availability of features to classify texts. Thus, the approach I propose for Swiss German is based on character ngrams, which not only avoids the problem of a lack of available NLP tools, but – in addition to being a proven successful feature for authorship attribution – allows the capturing of orthographical idiosyncrasies. It thus allows the exploitation of Swiss German's lack of standardised spelling rules, turning the challenge that Swiss German presents as non-standard data into an advantage. Different lengths of n-grams as features of a Na¨ıve Bayes classifier combined with varying sizes of training and test corpora were tested, and 6- and 7-grams were found to faultlessly identify authors for all combinations considered. The number of distinctive n-grams in an author's data set was found to be a determining factor for the classifier's success, highlighting the benefits of exploiting Swiss German's non-standard nature for authorship identification.

Originally published at:
Oppliger, Rahel (2016). Automatic authorship attribution based on character n-grams in Swiss German. In: KONVENS 2016, Bochum, 19 September 2016 - 21 September 2016. Universitätsverlag Ruhr-Universität Bochum, 177-185.

Automatic authorship attribution based on character n-grams in Swiss German

Rahel Oppliger
University of Zurich
rahel.oppliger@uzh.ch

Abstract

Automatic authorship attribution aims to train computers to identify the author of a disputed text based on idiolectal language features. When confronted with non-standard data – in the present study Swiss German instant messages – language-specific NLP toolkits are often unavailable, limiting the availability of features to classify texts. Thus, the approach I propose for Swiss German is based on character n-grams, which not only avoids the problem of a lack of available NLP tools, but – in addition to being a proven successful feature for authorship attribution – allows the capturing of orthographical idiosyncrasies. It thus allows the exploitation of Swiss German's lack of standardised spelling rules, turning the challenge that Swiss German presents as non-standard data into an advantage. Different lengths of n-grams as features of a Naïve Bayes classifier combined with varying sizes of training and test corpora were tested, and 6- and 7-grams were found to faultlessly identify authors for all combinations considered. The number of distinctive n-grams in an author's data set was found to be a determining factor for the classifier's success, highlighting the benefits of exploiting Swiss German's non-standard nature for authorship identification.

1 Introduction

Identifying the authors of texts purely based on stylometric evidence has been of interest to linguists since the 19th century, when Augustus DeMorgan suggested that authors can be identified according to the average word length in their texts, an idea taken up by Mendenhall (1887, p. 237). Since these early explorations, which have since been discovered to lack sufficient empirical foundation (Holmes, p. 88), various measures have been employed to determine authorship, ranging from sentence length, over vocabulary peculiarities, to the use of particular syntactic structures. Such methods have found application not only in determining the idiolectal styles of authors, but they have prominently been used to determine authorship of the Federalist Papers (Mosteller and Wallace, 1964), and they are predominantly employed in forensic linguistics – to aid in criminal cases that contain evidence of disputed authorship (Olsson and Luchjenbroers, 2013, pp. 7-9).

Over the last decades, large efforts have been undertaken to automate the process of authorship attribution, as well as to render it more empirically founded. Automatic authorship attribution, as this method has been termed, aims to determine the author of a disputed text reliably by identifying features indicative of the author's writing style utilising a corpus of known texts.

Since this new attribution paradigm relies solely on computers, the availability of a variety of natural language processing (NLP) tools directly influences the availability of features. The tools needed to process language for feature extraction range from tokenisers, over part-of-speech taggers, to syntactic parsers and semantic annotation tools. However, such extensive NLP toolkits are only available for the major languages that are regularly subjected to computational linguistic research.

As this study will be focused on Swiss German, the language independence of the method is vital considering this language is vastly under-researched, especially with regards to NLP. The difficulties of processing Swiss German lie predominantly in the fact that it is not a standardised language, and only recently have there been efforts to develop NLP tools for it. Hollenstein and Aepli (2014) have presented first steps towards develop-

ing a part-of-speech tagger for Swiss German, but more progress needs to be made until such a tool is fully dependable. Thus, the challenge of the present study is finding a reliable, standardised method for such non-standard data; in an attempt to achieve the best possible result despite the restrictions of software availability for Swiss German language data, the approach presented in the following will be based on character n-grams – a selection justified in Section 2.

As much as authorship attribution within a non-standard language like Swiss German poses problems in terms of software availability, there may also be peculiarities of a non-standard language to be exploited: in this study, the fact that Swiss German does not have standardised spelling and thus encourages individual spelling styles will be used to distinguish authors. The hypothesis I propose is that based on character n-grams that are able to capture the idiolectal orthography in written Swiss German, automatic authorship attribution is possible.

In the following, I will first review the literature on Swiss German orthography and n-gram-based automatic authorship attribution – justifying the selection of n-grams as a stylometric measure for my data – followed by a discussion of the processing necessary to train a successful classifier. I will analyse the group conversation of four authors on the instant messaging app WhatsApp, selecting a portion of the data to function as a training corpus from which to extract features to be used to automatically attribute authors to the messages in the second portion. This attribution process is carried out by a Naïve Bayes classifier that is trained on a feature set of character n-grams. Finally, I conclude with a discussion of how distinctive features interact with the success of the n-gram classifier.

2 Literature review

2.1 Feature selection

As part of the paradigm shift to computer-based methods in authorship attribution discussed in the introduction, the research community's aim has been to find features that are both easy to extract automatically and are maximally indicative of an author's written idiolect. Stamatatos (2009, pp. 540-544) presents a selection of such stylometric features that have been applied in authorship attribution, including lexical, character, syntactic, and semantic features. As he describes, the depth of text analysis and thus the complexity of NLP systems differs vastly from feature to feature: for instance, a lexical aspect such as word frequencies requires only a tokenised text, whereas the pre-processing needed to allow for the extraction of sentence and phrase structure features is extensive.

In an early overview of computer-based approaches, Holmes (1994) lists various features using elaborate pre-processing to differentiate authors: he questions the suitability of features such as word-length, sentence-length, and word frequencies, but he argues that approaches combining multiple features show particular promise. Such methods that incorporate various features are arguably able to capture an individual's idiolectic style more comprehensively. Over ten years later, Grieve (2007, pp. 266-67) reaches a similar conclusion: in his analysis of different approaches to automatic authorship attribution, he finds that by combining measurements, test accuracies above 93% are achieved for four possible authors – the number of authors considered in the present study – and for two authors, the accuracy is as high as 97%.

While the state of the art automatic authorship attribution approach is one that combines a variety of features, from a processing point of view it is preferable to focus on less features. The best-performing individual algorithm that Grieve (2007) tested for is one that distinguishes individuals according to how often they use various punctuation marks relative to the number of words in their texts. The success rate of this approach is 95% for two possible authors and 89% for four.

Both the construction of a word and punctuation profile and the combination approach require at least the availability of a tokeniser in the former case (to determine the number of words in a text) and a variety of NLP tools depending on the features being combined in the latter. Thus, an alternative approach must be found for non-standard data that lacks reliable or fully automatic tools. Houvardas and Stamatatos (2006, p. 78) identify language independence as one of the major advantages of using character n-grams for authorship attribution; since n-grams simply consist of consecutive strings of characters, no pre-processing is needed.

In addition to the practical advantages of extracting character n-grams, they have also been proven to perform very well in a variety of studies. Grieve (2007) examines thirty-nine different methods of

quantitative authorship attribution, and in applying them all to the same data set he finds that n-grams – particularly bi-, tri-, and 4-grams – are only surpassed in accuracy by the aforementioned feature-combining approaches and word and punctuation profiles. Specifically, he finds that both bi- and trigrams achieve 94% and 88% accuracy for two and four possible authors respectively, while 4-grams perform slightly worse at 93% and 85%.

While Kešelj et al. (2003) confirm the high success rates of the character n-gram approach, they find larger n-gram sizes to perform better: using sequences of 4 to 8 characters produces the best outcome. Yet more strikingly, 6- and 7-grams achieve 100% accuracy in distinguishing seven authors for feature profile sizes of 500 to 3,000 features. It has to be considered, however, that the pool of seven possible authors that Kešelj et al. distinguish between is quite varied: it ranges from 16th century Shakespeare to 19th century Lewis Carroll. In such a diverse pool of authors, the language samples considered can be expected to vary not only in idiolect but also rather drastically with respect to the historical period they were produced in, and the topic they are written about. The 100% accuracies that this study presents thus have to be regarded with caution. Nevertheless, it is of interest that they observe that 6- and 7-grams outperform other n-gram sizes.

The circumstances that are encountered in authorship attribution are often challenging: many authors may have to be considered, and the amount of available data may be very limited. The suitability of the n-gram approach in such difficult cases is further attested in Houvardas and Stamatatos (2006), who find that character n-grams work particularly well for multiple authors (p. 78). Moreover, as n-grams make it possible to extract a large number of features from even short texts (Layton et al., 2012, p. 299), they are well suited to registers usually producing brief messages, such as the instant messages considered here.

The approach based on character n-grams has also proven to be successful for social media data. In his application of a variety of features to the authorship disambiguation in Dutch tweets, van Halteren (submitted) finds that trigrams perform better than lexical features. He finds the success of trigrams to be particularly apparent in Twitter data – a data type that shares characteristics like short length of messages and a low degree of formality with instant messages. One potential reason he provides for the success of the feature is its ability to capture many characteristics of a tweet, such as capitalisation, spelling variation, user mentions, or URLs.

This success of n-grams in the classification of short computer-mediated texts may thus be attributed to their ability to capture different elements of language. Layton et al. (p. 298) argue that character-level n-grams are able to represent peculiarities on every level of language, whether the information be morphological, lexical, orthographical, or syntactical. In particular the benefits of n-grams being able to capture information on orthography will be discussed in the following section.

2.2 Swiss German's lack of standardised orthography

To further justify the selection of character n-grams as the ideal feature to differentiate between authors in Swiss German, we must consider how the language is written. Swiss German does not exhibit standardised spelling rules: every user of the language develops an individual set of spelling conventions. Ruef and Ueberwasser (2013, pp. 61-62) attribute this lack of uniform orthography to the large diversity in regional dialects, as well as to the low number of texts being written in this predominantly spoken language – Standard German is used for written communication in Switzerland, presenting a model case of diglossia.

In a corpus of Swiss text messages, sms4science, Ueberwasser (2013, p. 8) finds that almost two thirds of the German messages were in fact written in Swiss German dialect, thus rendering text messages a register where Swiss German appears in a written form. However, Ruef and Ueberwasser (2013) report that despite a growth of written Swiss German due to its uses in computer-mediate communication, virtually no standardisation has taken place, which may be partly caused by the register being largely informal, as official communication still uses Standard German.

The data set used in this study is indeed taken from a very informal context, namely a instant messaging conversation between friends; thus, the use of Swiss German is favoured. However, the dialectal heterogeneity that Ruef and Ueberwasser (2013) cite as a cause for lacking spelling norms is not a large issue in the present data set: three out

of four participants grew up in the same town, and although there are slight differences in the participants' spoken dialects, they are overall very similar. For such a homogeneous group, Scherrer and Rambow (2010, p. 98) voice concerns whether character n-grams are able to distinguish between very similar dialects. Yet, the hypothesis of this study is based on my observation that even between the four participants in this group chat – who all talk very similarly – there are considerable differences in spelling.

Contrary to Scherrer and Rambow's (2010) concerns, I argue that it is precisely with n-grams that we can successfully exploit the Swiss German particularity of exhibiting large variety in orthographic idiolects for the purpose of automatic authorship attribution. This non-standard feature of Swiss German lends itself to be captured in n-grams, and in the following, I will illustrate the degree to which spelling differs among individuals even of the same or a very similar dialectal background.

English / German		**P. 1**	**P. 2**	**P. 3**	**P. 4**
write, text / schreiben	schriebe	22	0	6	6
	schriibe	0	0	0	22
	schribe	1	10	0	0
Friday / Freitag	Fritig	2	0	0	4
	fritig	17	6	11	0
	Friitig	0	0	0	14
not / nicht	nöd	63	152	115	165
	ned	71	0	0	1
then / dann	denn	23	105	21	235
	den	127	1	77	0
now / jetzt	jetzt	5	12	2	136
	ez	52	0	79	0
	ezt	0	43	0	0

Table 1: Frequencies of orthographic alternatives in the WhatsApp training corpus.

Table 1 shows a selection of such spelling differences, presenting inconsistencies in spelling both between and within speakers. While Ueberwasser (2013, p. 20) suggests that an individual's spelling is often consistent, Table 1 illustrates that this certainly does not always apply. For example, in spelling 'now/jetzt', the four participants each prefer one of three spelling variants: participants 1 and 3 favour *ez* very strongly, while participant 2 prefers *ezt* and participant 4 *jetzt*. Participant 2 moreover exhibits considerable variation, using *jetzt* in approximately a fifth of instances. Showing even more intra-speaker variation, participant 1 uses both *nöd* and *ned* at almost equal frequency. However, in many cases, speakers indeed show a tendency to use a specific variant, as exemplified by the other participants' clearly preferred use of *nöd* over *ned*.

Furthermore, as Ueberwasser also notes (2003, p. 20), one speaker may be consistent in representing identical sounds with the same letter or letter combination. This phenomenon can be observed in the spellings for the mid-word vowel [i:] in 'write/text' and 'Friday': participant 4 prefers to represent the long vowel as /ii/, whereas participant 2 favours /i/. At the same time, however, participants 1 and 3 do not follow this pattern, using /ie/ in 'write/text' and /i/ in 'Friday'. These differences in spelling as well as potential consistencies in how individuals choose to represent certain sounds can be captured by character n-grams. Additionally, the distribution of how often the variants are used by each author is also represented in the n-gram feature profiles. In their ability to incorporate these particularities of non-standardised orthography, character n-grams form the ideal basis for an authorship attribution classifier for Swiss German. To sum up, this approach not only avoids the difficulties in working with this type of non-standard data by requiring no pre-processing, but it crucially exploits the data's idiosyncrasies. In the following, the extraction of the n-gram features and their use in a Naïve Bayes classifier will be discussed.

3 Data and method

3.1 Data

The data used in this study was obtained from a group chat on the instant messaging app WhatsApp between four Swiss females, aged 20 to 24 at the time of production. All participants have given their consent for me to use the data in this study. The four participants all share a similar spoken dialect. The conversation is conducted in Swiss German with occasional occurrences of English, Standard German, and French. The training corpus consists of 5,141 messages, varying in length and with different participants having contributed to various degrees; the exact size of the training (TR) and test (TE) corpora is presented in Table 2.

		P. 1	P. 2	P. 3	P. 4
TR	*msg.*	1,069	1,384	1,384	1,443
	char.	53,995	45,255	47,157	85,489
TE	*msg.*	330	405	393	285
	char.	19,271	17,016	13,571	19,186

Table 2: Size of the training and test corpora, in messages and characters per participant.

The split into training and test corpora was made at a specific point in time, resulting in the uneven sized corpora. This choice was motivated by an aim to control for the influence of topic of conversation; by splitting the data at the same point in the conversation for every participant, similar topics should be located in the training and test corpora of all participants. However, it is worth noting that the varying sizes of training material per participant may have an influence on the performance of the classifier. The test corpus that was used was a smaller part of the same conversation; roughly, the test corpus for each participant is 15-25% the size of the respective training corpus.

While the general Machine Learning principle of 'more data is more' applies to this task, too, Layckx and Daelemans (2010) set out to define the desired data size. They suggest that the ideal training set size lies above 10,000 words per author, but they acknowledge that with the use of n-grams, satisfactory results can be obtained on much smaller data sets (p. 53). How well a classifier based on character n-grams performs on data sizes below that desirable threshold will be explored in the present study. Namely, in addition to training and testing the classifier on the full data set, it is trained on half the training data size and tested on half, a fifth, and a tenth of the testing data size.

3.2 Method

In Section 2, I outlined the practicality and proven success of character n-grams as a feature for authorship attribution. The task is then to extract n-grams from the WhatsApp data – the conversations can be downloaded in the app as a plain text file. I extracted the messages for each participant and created n-grams of variable length for all messages, storing both the n-grams and how frequently they occur for every participant in feature dictionaries.

As the data is taken from an informal computer-mediated register – specifically instant messaging – we can observe an extensive use of emojis. Since these form a potentially defining part of an author's idiolect, they were included as part of the n-grams.

After the collection of n-grams of varying lengths, the resulting feature dictionaries that represent the language of each participant are used to train a classifier, specifically a Naïve Bayes classifier. Juola (2006, p. 285) cites the relative ease of training as one of the chief advantages of Naïve Bayes classifiers. In fact, Bird et al.'s (2009) Natural Language Toolkit (NLTK) contains a module that I use in this study to train a Naïve Bayes classifier and apply it to data.

In order to determine how well the classifier copes with different amounts of data, I test the classifier on a number of combinations of training and test data; I aim to determine whether it can still provide accurate results with lower amounts of data. Additionally, I attempt to add my results to the studies described in Section 2.1 that have sought to determine what size n-grams deliver the best results, thus testing bigrams to 10-grams, as Kešelj et al. (2003) did.

4 Results and discussion

4.1 Naïve Bayes classifier performance

I trained and tested Naïve Bayes classifiers on a number of training and test data set combinations, noting simply how many of the authors were correctly identified for each n-gram size and data size combination. These results are presented in Table 3, with the numbers in each instance referring to how many of the four authors were correctly identified by the classifier. It is evident that the Naïve Bayes classifier overall performs above chance for all sizes of training and test corpora examined here. In fact, with a vast amount of data available to both train and test the classifier on, i.e. the full training and test set, the performance is near faultless, with only bigrams and 4-grams failing to deliver fully correct results.

However, perhaps more interestingly, certain sizes of n-grams appear to produce wholly accurate results for all sizes of data sets tested: classifiers trained on 6-grams and 7-grams succeed in identifying the correct authors in all categories of data size, while the 8-gram classifier only decides incorrectly for one author in one category. These results match the findings of Kešelj et al. (2003), who also find 6- and 7-grams to be the most effective. The results of my study thus support their argument that these length n-grams are most indicative of individual

	Full training Full test	0.5 training 0.5 test	Full training 0.2 test	Full training 0.1 test	0.5 training 0.2 test	0.5 training 0.1 test
2-g	3	3	0	0	0	1
3-g	4	4	1	0	1	0
4-g	3	3	1	1	2	1
5-g	4	4	3	2	4	4
6-g	4	4	4	4	4	4
7-g	4	4	4	4	4	4
8-g	4	3	4	4	4	4
9-g	4	3	3	4	4	4
10-g	4	3	3	4	4	4

Table 3: Naïve Bayes classifier results, split by n-gram size and size of training and test set.

writing style.

More faulty performances can be seen from the classifiers trained on bigrams, trigrams, and 4-grams, where the classifiers perform at, or below, chance level when trained and tested on less data. Thus, Grieve's (2007) findings that short n-grams perform best could not be confirmed. However, it has to be noted that text type may play a considerable role in what size n-gram is most successful – the results presented here should therefore be regarded as potentially being particular to the instant messaging register and their transferability to other text types considered with caution. Nevertheless, it can be said that – with the right selection of n-gram size and sufficient data – character-based classifiers work very well in determining authorship in Swiss German instant messages.

Although the classifier evidently has trouble when trained and tested on shorter n-grams, namely bi- to 4-grams, it has to be noted that when the full size and half size training and test sets are used, the performance for these size n-grams is still well above chance level. Only when either the test or training corpus are substantially smaller does the classifier struggle. In the following, I will attempt to uncover the source of this problem.

4.2 Sparse data problem with short n-grams

As an explanation for the comparatively bad performance of shorter n-grams (bi- to 4-grams), I suggest that the issue in the present experiment is one of sparse data. A lack of sufficient training or test data has frequently been identified as one of the main problems Machine Learning approaches to authorship attribution face in forensic linguistic cases (Coulthard, 2004, p. 432; Totty et al., 1987, pp. 16-17). However, it is important to note that the sparsity of the data does not simply relate to the number of features, but to the number of distinctive features, as will be outlined in the following.

The shorter n-grams' success in the larger data sets may be attributed to their ability to capture the writer's language behaviour on a level that allows the classifier to compute a language profile for them. With less data, the creation of such a profile is seemingly not possible for the classifier, as simply not enough of the individual's habitual language behaviours may be present. Moreover, the profiles might be too similar as they are based more on frequent words and character sequences in the language rather than the individual's language choices. Longer n-grams, on the other hand, may be able to capture habitual language features even within a small amount of data, as they are less likely to produce identical features for all participants but will rather find a sufficient number of distinctive features.

	Part. 1	**Part. 2**	**Part. 3**	**Part. 4**
$2-gram$	11	7	2	18
$6-gram$	79	52	102	218

Table 4: Number of distinctive 2-grams and 6-grams found by the classifier for the half-sized training set and tenth-sized test set.

To illustrate this type of sparse data problem, I compare bigrams and 6-grams – the best and worst-performing n-grams – within the smallest data set in this study. In order for the classifier to be effective, distinctive features have to be found; a feature is distinctive if it appears both in the training and test corpus of one author, but not in the training corpora of the other authors. Table 4 shows that 6-grams provide far more distinctive features than bigrams. An examination of the distinctive bigrams reveals that they predominantly include emojis. To

sum up, the sparse data problem in n-gram classification has to be considered not at the level of how many features are available to train the classifier with, but how many of those features are distinctive.

4.3 Distinctive features

Following from the hypothesis that the success of an n-gram classifier is dependent on how many distinctive n-grams are available for each author, I now aim to illustrate the connection between number of distinctive n-grams and performance of the Naïve Bayes classifier.

According to my hypothesis, we would expect the n-gram sizes that produce the best results in the automatic authorship attribution task to produce the most distinctive n-grams. And indeed, as is shown in Figure 1 below, 6-grams exhibit the most distinctive features for three out of four authors, with participant 4 producing more unique 5-grams.

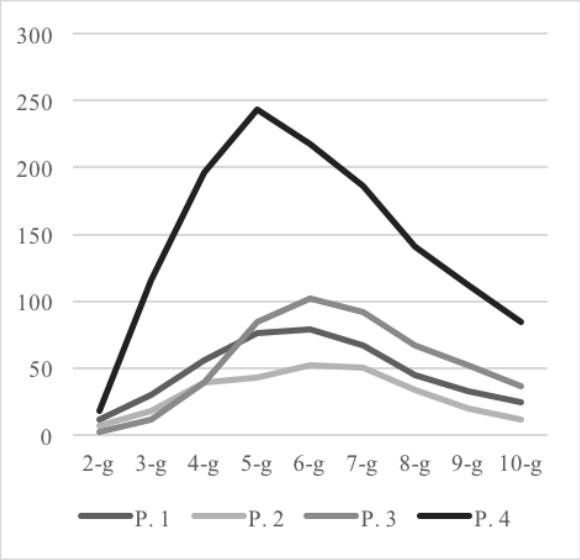

Figure 1: Distinctive n-grams for different n-gram lengths.

The n-gram sizes in Figure 1 reveal that the distribution for distinctive features peaks around 5- and 6-grams. As can be expected, participant 4, who provides the most training material, and thus allows for the extraction of more features, produces the most distinctive features for every n-gram size. However, the larger size of the training corpus shows the most benefits around its peak, while the very short and very long n-grams show similarly bad performances as with less data. I conclude from this case study that even with more training data available, the most efficient n-gram size, at least for this particular register, will be in the range of 5- to 8-grams, agreeing with Kešelj et al. (2003).

4.4 Distinctive spelling features

To illustrate how idiolectic spelling is reflected in the 6-grams that are found to be distinctive features, I will now take a look at participant 1's distinctive n-grams. A closer investigation of this participant's 79 distinctive 6-grams reveals that 23 involve spellings that are either highly indicative of this author or at least often used by her. Looking back at Table 1, where I demonstrated that participant 1 is the only author within this group to regularly – indeed over half of the time – use 'ned' to express *not/nicht*, it is perhaps not surprising that eight of the orthographically distinctive 6-grams contain this idiolectic spelling.

A further three distinctive sequences contain the lemma DERFEN, meaning *can/dürfen*, which is habitually spelled with a second vowel /e/ by participants 1 and 4, while participants 2 and 3 represent this vowel with /ö/. The /e/ vs. /ö/ distinction here is a similar one to that between 'ned' and 'nöd', and it is of interest to remark that participant 1 chooses the option /e/ in both cases. This observation supports Ueberwasser's (2013) hypothesis that identical sounds are often habitually represented by identical grapheme sequences.

Overall, consistent idiolectic spelling choices such as the aforementioned are here shown to be indicative of authorship, especially if they are characteristic of only the specific author. N-grams similar in length to the 6-grams investigated here are successful in capturing these orthographical idiosyncrasies. In this way, this characteristic orthographical freedom in Swiss German can be exploited as an effective feature for automatic authorship attribution.

5 Conclusion

In this paper, I have demonstrated that n-gram-based Naïve Bayes classifiers are successful in identifying authorship within four Swiss German speakers' instant messages. The outcome of this study leads me to conclude that character-based authorship attribution in Swiss German is a promising method, even for such small data sets as instant messages provide. I have suggested that the success of n-grams as features for identifying authorship in Swiss German may be amplified by the language's lack of standardised orthography,

encouraging each individual to develop their own spelling habits, which in turn may lead to a greater number of distinctive n-grams for the classifier to base its authorship attribution on.

Naïve Bayes classifiers based on different length n-grams delivered promising results in these authorship application tests, particularly for 6- and 7-grams, where perfect results for all data sets were achieved. These tests also revealed that the success of this method of authorship attribution lies in the classifier being able to find distinctive features within the data, creating a sparse data problem for shorter n-grams which fail to produce such distinctive features. Therefore, 5- to 7-grams proved to be the most suitable for the task, as they provide a sufficient number of distinctive features even within smaller data sets. Indeed, the number of distinctive features for any given n-gram size was found to correlate with how well the classifier performs.

While this study has presented promising results for character n-gram classifiers for automatic authorship attribution in Swiss German, further tests with a larger amount of authors and data will have to be undertaken in order to ensure the method's validity. Furthermore, rates of success will have to be tested more rigorously, particularly for forensic linguistic application.

Perhaps the most valuable finding of this study is that the non-standard nature of data is not merely a challenge to overcome, but that the particularities of a non-standard data set can be exploited. In this case, I have shown that Swiss German's characteristic lack of spelling rules – causing idiolectal orthography among its users – presents an opportunity to use this trait as an effective feature for automatic authorship attribution.

References

Steven Bird, Ewan Klein, and Edward Loper. (2009). *Natural Language Processing with Python*. Sabastopol, CA: O'Reilly Media, Inc.

Malcolm Coulthard. (2004). *Author identification, idiolect, and linguistic uniqueness. Applied Linguistics, 25*(4), 431-447.

Jack Grieve. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing, 22*(3), 251-270.

Hans van Halteren. (submitted). Large scale authorship recognition on productive Dutch-speaking Twitter users.

Nora Hollenstein and Noëmi Aepli. (2014). Compilation of a Swiss German dialect corpus and its application to PoS tagging. Paper presented at the *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, (pp. 85-94).

David I. Holmes. (1994). Authorship attribution. *Computers and the Humanities, 28*(2), 87-106.

John Houvardas and Efstathios Stamatatos. (2006). N-gram feature selection for authorship identification. Paper presented at the *Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, Applications*, (pp. 77-86).

Patrick Juola. (2006). Authorship Attribution. *Foundations and Trends in Information Retrieval, 1*(3), 233-334.

Vlada Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. (2003). N-gram-based author profiles for authorship attribution. Paper presented at the *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING*, (pp. 255-264).

Robert Layton, Paul Watters, and Richard Dazeley. (2012). Recentred local profiles for authorship attribution. *Natural Language Engineering, 18*(3), 293-312.

Kim Luyckx and Walter Daelemans. (2010). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing, 26*(1), 35-55.

Thomas C. Mendenhall. (1887). The characteristic curves of composition. *Science, 9*(214), 237-249.

Frederick Mosteller and David L. Wallace. (1964). *Inference and Disputed Authorship: The Federalist.* Reading, MA: Addison-Wesley.

John Olsson and June Luchjenbroers. (2013). *Forensic linguistics*. London: A&C Black.

Fuchun Peng, Dale Schuurmans, Vlado Kešelj, and Shaojun Wang. (2003). Language independent authorship attribution using character level language models. Paper presented at the *Proceedings of the tenth Conference of European Chapter of the Association for Computational Linguistics*, (pp. 267-274).

Beni Ruef and Simone Ueberwasser. (2013). The Taming of a Dialect: Interlinear Glossing of Swiss German Text Messages. In C. M. Bongartz and C. M. Riehl (Eds.), *Non-standard Data Sources in Corpus-Base Research*, (pp. 61-68). Aachen: Shaker.

Yves Scherrer and Owen Rambow. (2010). Natural language processing for the Swiss German dialect area. Paper presented at the *Proceedings of KONVENS 2010*, (pp. 93-102).

Efstathios Stamatatos. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology, 60*(3), 538-556.

R. N. Totty, R. A. Hardcastle, and J. Pearson. (1987). Forensic linguistics: The determination of authorship from habits of style. *Journal of the Forensic Science Society, 27*(1), 13-28.

Simone Ueberwasser. (2013). Non-standard Data in Swiss Text Messages with a Special Focus on Dialectal Forms. In C. M. Bongartz and C. M. Riehl (Eds.), *Non-standard Data Sources in Corpus-Based Research*, (pp. 7-24). Aachen: Shaker.